The quest for absolute abundance: the use of internal standards for DNA-based microbial and community ecology

Joshua G. Harrison[1,2], W. John Calder[1], Bryan Shuman[1], and C. Alex Buerkle[1]

[1]University of Wyoming,
Laramie, WY 82071, USA

[2]*Corresponding author*: Joshua G. Harrison
1000 E. University Ave.
Department of Botany, 3165
University of Wyoming
Laramie, WY 82071, USA
joshua.harrison@uwyo.edu
Fax: 307-766-2851

Keywords: microbiome, internal standard, spike-in, compositional data, absolute abundances, relative abundances, microbial ecology, metabarcoding

Running title: Internal standards

Abstract

To characterize microbiomes and other ecological assemblages, ecologists routinely sequence and compare loci that differ among focal taxa. Counts of these sequences convey information regarding the occurrence and relative abundances of taxa, but provide no direct measure of their absolute abundances, due to the technical limitations of the sequencing process. The relative abundances in compositional data are inherently constrained and difficult to interpret. The incorporation of internal standards (ISDs; colloquially referred to as “spike-ins”) into DNA pools can ameliorate the problems posed by relative abundance data and allow absolute abundances to be approximated. Unfortunately, many laboratory and sampling biases cause ISDs to underperform or fail. Here, we discuss how careful deployment of ISDs can avoid these complications and be an integral component of well-designed studies seeking to characterize ecological assemblages via sequencing of DNA.

Introduction

Ecological assemblages, particularly microbiomes, are routinely characterized by DNA sequencing of marker loci, which are typically short and are chosen because they vary among focal taxa (Caporaso et al. 2012; Carini 2019; Goodrich et al. 2014)—portions of the ribosomal RNA operon are particularly popular markers. Characterizing assemblages in this way is referred to as metabarcoding (Schmidt et al. 2013; Taberlet et al. 2012). Qualitative differences in the sequences obtained from a metabarcoding study can be used to generate hypotheses regarding the types of organisms present in an assemblage, but understanding the abundances of each of these organisms from sequence data alone has proven extremely challenging. This is because sequencing methods yield a platform-specific amount of data (i.e., reads), which are then parsed among samples and molecules within each sample. Thus, metabarcoding can only provide direct knowledge of the relative abundances of organisms, not their absolute abundances. The same technical challenges apply when performing other types of sequencing, including shotgun metagenomics and transcriptomics (Chen et al. 2016), thus relative abundance data are ubiquitous across molecular ecology disciplines.

Analyzing relative abundances is challenging for several reasons. First, biological insights often depend on knowledge of absolute abundances. For instance, in a study of the faecal microbiome of patients with Crohn's disease, absolute abundance data (obtained through flow cytometry) revealed that bacterial load was associated with disease phenotype (Vandeputte et al. 2017)—an unobtainable result when using relative abundance data. More generally, dramatically different results were obtained from analyses of absolute versus relative abundance data. For example, the use of absolute abundance data led to detection of 76 covarying microbial genera, compared to detection of only 10 covarying genera when using relative abundance information. Relative abundance data were misleading about microbial richness, rank abundances, and associations of specific taxa with disease phenotype—thus demonstrating that relative abundance data are unsuitable for addressing many biological questions (for a similar example see Stämmler et al. 2016).

The problems associated with relative abundances largely stem from their compositional nature (Aitchison 1982), that is, as one taxon increases within a sample, it does so relative to some other taxon (or taxa) that must decrease (Fig. 1). For over a hundred years, mathematicians have been aware of the numerous problems associated with the analysis of compositional data (Pearson 1897). Indeed, many of the standard multivariate tools useful for community ecology are inappropriate for compositional data (see Gloor et al. 2017; Jackson 1997). Several sub-fields of ecology have developed rich literatures about these complications (Jackson 1997) with associated disciplinary names for the challenges of compositionality, such as the 'Fagerlind effect' (i.e. a term used in paleoecology to refer to the problems inherent to the analysis of compositional pollen data), which complicates cross-disciplinary transfer of relevant information (Davis 1963; Fagerlind 1952; Prentice and Webb 1986). Nevertheless, acknowledgement of the constraints imposed by compositional data is becoming more commonplace among ecologists, particularly those characterizing microbiomes via sequencing data (Gloor and Reid 2016; Weiss et al. 2017). Still, many studies do not adequately confront the problem of compositionality and are hampered by the limitations of relative abundance data.

A variety of statistical transformations involving log ratios have been suggested to address the problems of compositionality, with perhaps the most common being the centered log ratio (clr) transformation (Aitchison 1982; Egozcue et al. 2003; Fernandes et al. 2014; Gloor et al. 2017). However, the benefits of the clr transformation are limited for high-dimensional, sparse data (data with many zeros, such as those describing assemblages with numerous rare taxa, which may not be observed at all in many samples), such as those characterizing microbial biodiversity. This is because logs of zero are undefined and thus, sparse data requires the addition of some constant to every element. The geometric mean of high dimensional, sparse data approaches this constant and thus ceases to provide a normalization benefit when used as a divisor (for more see Tsilimigras and Fodor 2016). Furthermore, the transformations alone do not allow for the conversion of relative abundance estimates to absolute abundances.

A promising solution to these problems is the incorporation of an internal standard (ISD) into the DNA sequencing process (Chen et al. 2016; Hossain et al. 2020; Jiang et al. 2011; Smets et al. 2016; Tourlousse et al. 2017; Zemb et al. 2020). Colloquially, this process is referred to as adding a “spike-in” of known quantity to samples. Similar approaches to spiking samples with an ISD have been applied in other disciplines seeking absolute abundances (e.g. paleoecology; Benninghoff 1962; Davis 1966; Davis and Deevey 1964; Giesecke and Fontana 2008). For high-throughput sequencing, the relevant ISD is a unique molecule (or cell, see below) that is added to all samples in a known absolute abundance (i.e., as measured in cells or moles). Through comparison to the ISD, the relative abundances of other sequenced features can be converted to units of absolute abundance (see below for an example; Fig. 1). ISDs are powerful tools that are rapidly gaining attention, particularly among microbial ecologists, but they are still not routinely used. As ISDs become regarded as critical components of a well-designed sequencing study (Chen et al. 2016; Jones et al. 2015), there is a need for understanding of the many commonly-encountered sampling scenarios and the laboratory biases that can undercut the efficacy of the standards. Here, we describe these considerations and suggest best practices for the design and use of ISDs. Much of our discussion relies on analogy to and examples from the microbial ecology literature, with specific application to metabarcoding, however our review is broadly relevant to characterization of absolute abundances of nucleic acids as required across sub-disciplines of molecular ecology using a variety of techniques (e.g., environmental DNA sequencing for metabarcoding of vertebrate taxa, metagenomics, qPCR, transcriptomics, etc.).

Is an ISD needed?

Prior to designing an ISD suitable for a particular study design, it is worth considering if an ISD is needed. For instance, if the sample can be homogenized to allow counting of target cells within an aliquot then an ISD will provide little additional benefit—though it could still act as a positive control and provide insight into technical variation. Counting cells may be possible for studies with few samples and can be accomplished through fluorescence microscopy (Amann and Fuchs 2008; Daims et al. 2001) or flow cytometry (Props et al. 2017a,b). For example, Vandeputte et al. (2017) used flow cytometry to count cells within a series of faecal samples and used these counts to transform 16S data from relative to actual abundances (also see Frossard et al. 2016). Such approaches hold great merit because many of the concerns with ISD efficacy that we describe below would be obviated by having a cell count in hand. Unfortunately, optimizing flow cytometry protocols for experimental conditions may be impractical for many researchers, particularly those studying microbial assemblages living inside tissues of a host organism (Doležel et al. 2007). Moreover, flow cytometry requires specialized equipment and skill, and can increase the logistical burden of a study more than the use of a spike-in ISD.

Quantitative PCR can also be used to estimate total copies of a genomic feature in a sample (e.g., copies of 16S), which can then be used to convert relative abundance estimates for each taxon to absolute abundances (Bonk et al. 2018; Dannemiller et al. 2014; Higuchi et al. 1993; Jian et al. 2020; Lou et al. 2018; Zhang et al. 2017). Droplet digital PCR (ddPCR; Hindson et al. 2011), is a promising tool for this approach because it provides heightened accuracy and throughput compared to conventional real-time qPCR; most importantly, it estimates abundances directly and does not rely on comparison to a quantitative standard (Baker 2012; Hindson et al. 2011; Kim et al. 2015; Morella et al. 2018). Barlow et al. (2020) recently used such an approach to demonstrate that absolute abundances of gut bacteria shifted in mice eating a ketogenic diet, and that relative abundances of particular taxa gave misleading results compared to absolute abundances. At the time of writing, ddPCR is currently more expensive than qPCR and also operates over a smaller dynamic range. The use of qPCR, via ddPCR or traditional techniques, is a simple, elegant approach to estimate absolute microbial abundances, however many of the pitfalls affecting ISDs can also affect this technique (Bonk et al. 2018). Moreover, while qPCR is relatively inexpensive, costs can mount when analyzing many thousands of samples and, therefore, the use of an ISD may save time and money for large-scale sequencing studies. The benefits and drawbacks of qPCR versus ISDs are poorly characterized, however, Stämmler et al. 2016 suggested that cellular ISDs outperformed qPCR for conversion of relative abundances to absolute abundances. These authors were studying the faecal microbiome and it is unclear if their findings translate to other sample types.

How does an internal standard work?

The potential benefit of ISDs is that they allow the conversion of relative abundances into absolute abundances. To see why this is desirable and why relative abundances in compositional data are problematic, consider a hypothetical comparison of two microbiome samples (Fig. 1). The first sample contains two equally-abundant microbial taxa and the second sample contains the same two taxa, but their relative abundances have shifted such that one is more abundant than the other. We could represent sequence data for these samples as vectors of proportions, with the first sample consisting of two equally abundant elements with proportions that sum to one $\vec{p}_1 = [0.5, 0.5]$. Whereas, the second sample has unequal elements, but the proportions also sum to one, e.g.: $\vec{p}_2 = [0.7, 0.3]$. The fact that both vectors must share the same sum (1 in this case) is referred to as the “constant sum constraint” of compositional data (Gloor et al. 2017) and is why neither of these vectors, nor the underlying sequence data, contain direct information regarding the absolute abundances of the microbial taxa being examined. For instance, it is impossible to know why, in sample two, the first microbe is greater in relative abundance compared to sample one. The difference could be due to the first taxon truly having a higher absolute abundance in sample two than in sample one. But it could also be due to a *decrease* in the second microbial taxon, or some combination of both possibilities, because the constant sum constraint of relative abundance data must be satisfied.

This conundrum can potentially be resolved if a known quantity of a third microbial taxon is added to each sample as an ISD (Fig. 1, panels g and h). Continuing with the previous example, we could include an ISD as the third element of each sample. After adding the *same* number of cells of the ISD to both microbial samples and repeating the sequencing process, one might obtain a proportion vector for sample one of: $\vec{p}_1 = [0.45, 0.45, 0.1]$, and for sample two of $\vec{p}_2 = [0.7, 0.25, 0.05]$ (the proportion taken by the ISD, the third number, could take any non-zero value). Because the same cell count of ISD was added to each sample, calculating the ratio of microbial relative abundances to the relative abundance of the ISD transforms the relative abundances making them proportional to absolute abundances, with units of the ISD (Fig. 2). In the example, on the scale of the ISD, the absolute abundances in sample one are $[4.5, 4.5, 1]$ and in sample two are $[14, 5, 1]$. Thus, for every unit of ISD observed there were 14 units of the first microbial taxon in sample two, but only 4.5 in sample one, indicating that the first microbial taxon is present at higher absolute abundance in sample two. The second microbial taxon also increased in abundance in sample two compared to sample one, but did not do so as much as the first taxon. Absolute abundances in units of the ISD can be scaled appropriately to other units by knowing the amount of standard that was added (the number of cells, or the number of moles of a DNA molecule).

If the log of the ratio between the ISD and each feature is taken then the aforementioned calculation becomes a case of the 'additive log ratio' (alr) transformation (Aitchison 1982). The alr is a popular transform in compositional data analysis and is expressed as:

$$alr(\vec{x}) = \vec{y} = \left[ln\frac{x_1}{x_D}; ...; ln\frac{x_{D-1}}{x_D} \right]$$

where $\vec{x}$ is a simplex with D components. The alr maps the simplex onto the real numbers, thus allowing multivariate statistics to be applied, so long as those statistics do not assume a preservation of relative distances among the elements of the transformed vector (see Aitchison and Egozcue 2005; Gloor et al. 2017; Quinn et al. 2018, 2019; Tsilimigras and Fodor 2016, for more). The choice of denominator in this transform is arbitrary. We mention the alr, and point the reader to aforementioned citations, to provide an avenue to explore the rich field of compositional data analysis, while noting that the primary benefit of ISD use is to sidestep the problems of compositionality.

What type of internal standard should be used?

Two main approaches exist for using ISDs in sequencing studies. The first involves adding a foreign molecule (or cell) to samples to be sequenced; we will refer to this method as a "spike-in" ISD. Alternatively, invariant features already present within samples can be used; we will refer to this type of ISD as an "inherent" ISD.

Researchers studying gene expression have long relied on inherent ISDs to facilitate comparison of transcription levels across samples (reviewed by Eisenberg and Levanon 2013; Thellin et al. 1999). Inherent ISDs are chosen from among those genes that contribute to the basic functioning of the cell (“housekeeping” genes) and are thus expected to be constitutively expressed. The idea is that these genes constantly produce the same number of transcripts, thus reads from them can be used as a baseline when comparing the expression levels of other genes among samples. Identifying housekeeping genes that are suitable for use as inherent ISDs is challenging and highly system-dependent because expressed genes differ among organisms and tissues, and the assumption of constitutive expression is often violated (Eisenberg and Levanon 2013; Jonge et al. 2007; Lun et al. 2017; Thellin et al. 1999; Tricarico et al. 2002). These drawbacks eliminate inherent ISDs from consideration for molecular community ecology—clearly, no taxon is expected to exist at identical abundances among habitats.

Molecular community ecologists thus must rely on spike-in ISDs. The development of spike-in ISDs has proven challenging, however, because the following assumptions must be satisfied: 1.) the ISD must behave similarly to template nucleic acids during laboratory practices, a characteristic referred to as “commutability” (Hardwick et al. 2017; Risso et al. 2014); and, 2.) there can be no chance that the ISD can be mistaken for a feature naturally occurring in samples. A third, practical consideration is deciding when the spike-in should be added during laboratory procedures and determining how much of it to add (as discussed below). Of these challenges, designing an ISD with sufficient commutability is the most daunting because ecological communities typically contain many taxa with vastly different traits—including variation in cell wall structure that influences cell lysability and thus DNA extraction yield. Similarly, even a pool of purified DNAs from various taxa will differ in primer affinity, sequence length, GC content, and so on, all of which can affect PCR performance (Bonk et al. 2018).

Two broad types of spike-in ISDs have been developed for metabarcoding: cellular ISDs and DNA ISDs. Cellular ISDs consist of adding cells of a foreign taxon to each sample, while DNA ISDs consist of DNA that has been extracted from an organism or synthesized. Both types of ISDs provide unique benefits for solving the commutability problem, but, unfortunately, both also have drawbacks, as we will discuss.

To our knowledge, cellular ISDs were the first to be used for metabarcoding (Jones et al. 2015; Stämmler et al. 2016); for example, in a seminal paper Stämmler et al. 2016 suggested using cells of several halophilic bacterial taxa and one bacterial taxon that occurs in the plant rhizosphere as ISDs for studies of the mammalian faecal microbiome. Because cellular ISDs were added prior to extraction, they allowed for measurement of variation in extraction yield among samples, at least to some extent. Indeed, since cells can drastically differ in amenability to DNA extraction (e.g., Gram positive versus Gram negative cells) and the sample matrix can also affect extraction performance, well-chosen cellular ISDs could potentially improve commutability for many studies.

The downsides to cellular ISDs are two-fold: first, choosing a cellular ISD can be challenging because it must have similar traits to focal organisms (so that behaves similarly to those organisms during extraction and PCR), be easily cultured (or available commercially), and cannot occur in the biological samples. Second, a non-clonal culture of a cellular ISD could possess copy number variation (CNV) in marker loci that must be measured and accounted for, else the ISD will not provide consistent and accurate absolute abundance estimates (Kembel et al. 2012). Even for clonally propagated ISDs, CNV for marker loci still must be determined to ensure accurate estimation of absolute abundances. For well-known taxa, estimates of CNV for marker loci could be obtained from published genomic resources (Langille et al. 2013; Perisin et al. 2016; Stoddard et al. 2015) or, for less studied taxa, quantitative PCR (qPCR) could be used to estimate copy number per cell. For those ecologists interested in non-microbial assemblages, determining suitable cellular ISDs is particularly challenging because culturing cells that are commutable with focal taxa may not be possible.

An alternative approach to cellular ISDs is the use of DNA molecules. Many microbial ecologists have advocated DNA ISDs, either in the form of extracted genomic DNA from organisms not likely to be present in samples or as synthetically designed molecules (Hardwick et al. 2016, 2018; Lin et al. 2019; Smets et al. 2016; Tkacz et al. 2018; Tourlousse et al. 2017; Venkataraman et al. 2018; Yang et al. 2018; Zemb et al. 2020). We suggest that synthetic sequences are superior to biologically-derived DNA for several reasons. First, and most obviously, there is no chance a synthetic sequence will occur naturally in samples, regardless of sample type. Second, reference DNA that is isolated from the genome could correspond to a variable number of genomic loci (CNV; as would actual cells; see above) and accounting for this potential variation among different isolates of a standard would require additional laboratory work, such as qPCR. Third, the nucleotide composition of an extracted DNA sequence is fixed and will likely only be commutable to a subset of focal taxa. By comparison, a synthetic ISD's DNA sequence can be specified such that it is comparable to the nucleotide composition of any organism (e.g., in length, GC content, repeat density, etc.) and thus could be tailored to fit the specific needs of a study.

The design of a synthetic ISD is fairly simple. The primary requirements are that the sequence cannot match any known organisms and is long enough that it will not be removed during PCR clean up (e.g., when using size selection to remove excess primer molecules). If a generic ISD is desired, then the sequence should minimize homopolymers and internal complementarity, have balanced GC content, and be approximately the same length as the focal metabarcoding locus. Alternatively, the sequence(s) could be designed to mimic focal taxa even if emulation could produce less than ideal sequence characteristics, thus potentially improving commutability during PCR and sequencing. After designing the ISD sequence, it must be bracketed by the preferred primer pair, with the complement of the forward primer at the beginning of the read and the uncomplemented reverse primer appended to the read (assuming single stranded synthesis). A variety of ISD designs are present in the literature (Table 1). Designs range from fully synthetic to hybrids between synthetic and biological sequences. For example, (Tourlousse et al. 2017) interject non-biological, synthetic sequences into the full-length 16S sequence of *Escherichia coli* and several other bacteria, thus allowing ISD sequences to be differentiated during analysis, but ensuring that they mimic many aspects of the 16S architecture. Hardwick et al. 2018 describe an elegant approach to ensure ISDs emulate focal taxa during laboratory preparation through preserving sequence composition characteristics (e.g., GC content, etc.). These researchers suggest simply reversing the portion of the genome of the focal taxon under consideration (e.g., the portion of the rRNA operon commonly used for molecular metabarcoding). The approach of Hardwick et al. 2018 was suggested for shotgun metagenomics. Notably, if such a technique is used for single-locus, metabarcoding, correct-sense primer sequences must be appended to the reversed sequence to ensure amplification.

Trade-offs exist with all ISDs such that a general statement regarding the superiority of any approach would be misleading. However, we suggest that actual microbial cells should be used as ISDs for studies involving samples that are likely to vary in nucleic extraction yield and for which certain focal taxa are known, such that a commutable ISD(s) could be chosen. We acknowledge that for many experimental designs commutable cellular ISD(s) could be difficult to choose. In such a situation, synthetic DNA ISDs could be simpler to use and thus preferable. Synthetic DNA ISDs could also be used for studies where samples are not likely to vary systematically in extraction performance (e.g., leaves from the same plant taxon; aliquots of similar soils). We do not advocate the use of extracted genomic DNA as an ISD unless CNV for focal loci is known.

The benefits of ISD mixtures

Regardless of whether a study design dictates the use of a cellular or synthetic ISD, researchers should consider the benefits of using a mixture of multiple ISDs as opposed to a single sequence or taxon. By adding a known amount of multiple ISDs to each sample, the failure of any one ISD to act as a true standard can be detected (Ji et al. 2020). For instance, if three ISDs were added to each sample in equal abundance and the relative abundance of the ISDs in the data obtained from the sequencer for a particular sample were 1:2:1, then it is clear that the second ISD was over-represented and should be omitted from consideration for that sample. Identification of a single malfunctioning standard is possible when using three (or more) standards, whereas if only two standards were used it would not be possible to determine which of the two ISDs had failed.

Another benefit of a mixture of ISDs is that it may lead to increased robustness to technical variation. For instance, Tourlousse et al. 2017 created 12 synthetic ISDs and reported that each responded slightly differently to laboratory practices. Accordingly, they reported an improvement in the accuracy of absolute abundance calculations when summing read counts across ISDs. The same result was reported by Stämmler et al. 2016, who used several cellular ISDs.

A final benefit of an ISD mixture is that sequences (or cells) emulating a variety of taxa can be included; thus, providing insight into the effects of laboratory practices across taxa akin to using a mock community as a positive control (Goodrich et al. 2014; Nguyen et al. 2014). Clearly, as ISD mixtures become more complex, they demand more sequencing depth—saying nothing of the time spent on their design. Until a sufficient breadth of ISD mixtures becomes commercially available, we suggest that researchers strike a balance between commutability and logistical cost by choosing a handful of sequences (or cells) that emulate those of focal taxa.

Considerations when deploying an ISD

The primary reason ISDs can fail to act as a standard is when the ratio of focal cells (or sequences) to the ISD shifts among samples in unexpected and unmeasured ways (Fig. 2, 3). A simple way this can happen is if there is unmeasured and unaccounted for variation among samples in input mass. To see why this is problematic, consider the situation in which two samples have identical microbial assemblages, but one sample has half the input mass of the other sample and therefore contains half as much DNA (Fig. 2c). If the same amount of ISD were added to each sample and normalization calculations performed as described above without accounting for sample mass differences, then it would appear as if microbial abundance was twice as high for one of the samples. While the two samples truly differ in microbial abundance, the difference is driven by differences in input mass among samples, not by differences in the microbial density in the source material. Consequently, laboratory methods typically involve standardization of the input mass of samples. However, imprecision in mass measurements made prior to nucleic acid extraction is rarely accounted for during data analysis and can add misleading variation to absolute abundance estimates. Problematic confounding could arise if sample mass were to differ systematically by substrate, experimental treatment, or among other batches. Fortunately, if input mass or volume varied among samples but was recorded, researchers can transform absolute abundances to absolute densities, on a scale of units of the ISD per unit of input mass (or volume).

A more insidious problem is when samples possess similar total masses but differ in the amount of target substrate present. For instance, if samples differ in hydration, then variation in the amount of water present could obscure differences in extractable mass among samples. Therefore, samples should be well dried prior to weighing and ISD incorporation. Variation in the amount of inorganic substrate present is particularly challenging for soil samples, which often differ in mineral composition, and hence density. In such cases, researchers should consider if volume is a more appropriate unit by which to standardize samples. The problem becomes amplified by comparisons across different substrates with fundamentally different characteristics and varying mixtures of potential microbial 'habitats' (e.g., comparisons across water, soil, and plants, or even different soils containing assemblages derived from communities within pore water, organic, and inorganic matter pools). Two soils could have identical water masses and contain the same microbial taxa, but varying soil matrices and associated microbial masses, which could alter the final homogenized samples if normalized by total volume or mass. In such cases, samples may require separation to better allow normalization of the target fraction (e.g., the organic portions of soil samples).

Time represented by the sample may also be important (e.g., duration of water filtration or sediment accumulation) because, all else being equal, more biological cells are likely contained within samples that encompass greater time and thus been subject to greater cellular deposition. Variation in the time captured by a sample could be particularly problematic when attempting to quantitatively compare assemblages via environmental DNA, such as when using cells in lake sediment to characterize aquatic invertebrate and vertebrate assemblages (Thomsen and Willerslev 2015; Turner et al. 2015).

ISD efficacy can also be undercut by variation in nucleic extraction performance among samples (Fig. 2c). For instance, if samples differ in physical toughness, such as what could be expected among tissue types of plants (i.e., stems versus leaves), more DNA will be obtained from samples with cells that are easier to lyse and the ratio of ISD to template DNA obtained will shift among samples, leading to inaccurate absolute abundance calculations. The same problem could occur if samples differ in the presence of compounds that inhibit extraction effectiveness (e.g., phenols in plants; Wilson 1997).

Variation in extraction yield is particularly difficult to measure for researchers interested in endosymbiotic microbial assemblages. This is because the recalcitrance of samples is defined by the traits of the host cells within and among which focal microbes reside (e.g., cell wall thickness can vary among plant taxa and tissue type) and a microbial cellular ISD will not emulate these traits. A possible solution for this problem is suggested through recent work by Karasov et al. (2019) who show that host-derived DNA can function as an inherent ISD when examining microbial symbiont assemblages. These researchers suggest estimation of microbial load as the ratio of host to bacterial reads obtained from shotgun metagenomic sequencing (also see Guo et al. 2019; Humphrey and Whiteman 2020; Karasov et al. 2018, 2019; Regalado et al. 2019). A possible benefit of this approach, as stated, is that metagenomic sequencing is a less biased way to estimate total host and bacterial load than amplicon sequencing.

Unfortunately, nucleic acid extraction methodology is not the only laboratory technique that can influence the effectiveness of an ISD. Compounds that can inhibit or facilitate PCR (Rossen et al. 1992; Wilson and Carroll 1997) may also cause problems. Consider the case when variation in amplification has occurred across samples that differ only in the presence of inhibiting or facilitating compounds (reviewed by Schrader et al. 2012). Assuming commutability, an ISD could account for these effects. However, Huggett et al. 2008 report variation in inhibition across PCR reactions. The drivers of this inhibition were unclear, but the authors suggested variation in amplicon GC content and primer melting point were two possible causes. Opel et al. 2010 reported similar sequence-specific inhibition and found that the mode of action varied markedly among compounds. These studies confirm that inhibitors can act in a sequence-specific way, which would undercut the commutability of ISDs for some portion of the amplicon pool they represent.

We are unaware of any studies or software that model the sequence qualities (e.g., length, GC content, etc.) that could lead to PCR inhibition in the presence of various compounds. We suggest that understanding the effect of PCR inhibitors on taxa of particular biological interest (e.g., important pathogens) and within oft-studied substances (e.g., blood, urine, tissues of model organisms) is a pressing need. Because of the looming issue of PCR inhibitors, we suggest that nucleic acid extraction protocols be preferred that consistently remove problematic compounds at the outset—a stated benefit of many commercially available extraction kits (e.g., the Qiagen PowerSoil kit removes humic acid; Mahmoudi et al. 2011). Similarly, we suggest that compounds known to block the action of inhibitors be considered as additions to PCR recipes (e.g., bovine serum albumin; Opel et al. 2010) and that modern polymerases (e.g., the Thermo-Scientific Phire and Phusion polymerases) be employed as they can bind to DNA more strongly than earlier commercialized versions of the polymerase enzyme (Flores et al. 2012; Videvall et al. 2017).

Given the many ways an ISD can fail, we suggest researchers incorporate several control measures into sequencing studies to ensure ISDs perform as expected. At the minimum, ISDs should be added to technical replicates of samples representative of the biological variation present. Upon sequencing, the ISD should capture approximately the same proportion of reads in each of these replicates. Secondly, as mentioned above, we advocate for using a mixture composed of at least three ISDs. Third, when using a new ISD, or using an ISD during sequencing of an unfamiliar substance, it is ideal to test for quantitative behavior through sequencing a dilution series; reads should increase proportionally to ISD concentration. Fourth, the possible confounding effects of inhibitors should be kept in mind, and, if possible, explored for the experimental system under consideration. Finally, we suggest that PCR cycles be kept to a minimum to avoid allowing PCR to continue until the stationary phase (Kelly et al. 2019).

At what laboratory step should an ISD be added?

To ensure spike-in ISDs perform properly, they must be added to samples at an appropriate time. Most authors advocate adding the ISD before nucleic acid extraction, and we concur (Jones et al. 2015; Smets et al. 2016; Tkacz et al. 2018; Tourlousse et al. 2017; Venkataraman et al. 2018; Zemb et al. 2020). This allows an ISD to capture variation in extraction performance (as mentioned above; Fig. 2b). Tkacz et al. (2018) added ISDs to soil samples both before and after DNA-extraction and report superior performance when ISDs were added before extraction. We note that measuring variation in extraction yield is best achieved via a cellular ISD that mimics traits of focal taxa (see above), however adding a DNA ISD to samples pre-extraction is also be beneficial (Zemb et al. 2020). The benefit of the latter approach arises because the abundance of the ISD in the sample would track the expected and potentially variable loss of DNA in extraction, such as would be caused by incomplete processing of all sample mass, variance during movement of supernatant and sample mass through the extraction protocol, or variable elution of nucleic acids from the solid-phase of columns used to isolate those acids.

If samples come from the same substrate and are thus not expected to behave differently during nucleic acid extraction, then an ISD could be added after extraction but prior to normalizing DNA concentrations for PCR (Fig. 2a). Though we acknowledge that adding an ISD at this step is less than ideal, given potentially unknown characteristics of samples that could have affected extraction yield. Notably, if an ISD is added after equimolar normalization of input DNA, then it will not be possible to accurately estimate absolute abundances in the original samples (Fig. 2d) because there will be no variation in the ISD among samples. However, even in this limited case, the ISD could still perform a useful role as a constant, positive control for PCR and sequencing.

How much ISD should be included in samples?

Choosing how much ISD to add to each sample can be challenging. Of course, it is important that the ISD be added in such quantity that it is detectable in all samples after sequencing, but it is also important to avoid adding so much ISD as to waste sequencing bandwidth. The majority of studies we considered showed expected quantitative behavior of ISDs throughout a wide range of input concentrations (e.g., Stämmler et al. 2016; Tourlousse et al. 2018), including quite low ISD input (~0.1% of the expected focal DNA mass present, see Smets et al. 2016). However, we acknowledge that for many substrates, homogenization of the sample prior to extraction is challenging and it is likely that some ISD will be bound up in unextracted material. Therefore, we suggest sacrificing some sequencing bandwidth to ensure the ISD is present in all samples. We suggest that 1–3% of the expected DNA yield is a reasonable target concentration for ISD addition (following Lin et al. 2019; Piwosz et al. 2018). We note that if extreme sequencing depth is employed, such as what can be obtained through the Illumina NovaSeq platform, it may be possible to use much less ISD and still achieve satisfactory results. We also suggest that a modeling approach to estimate proportions from count data for all sequenced features should allow much lower input of ISD than would estimation of proportions following rarefaction, because accurate estimates of proportions can be modeled given few observations (Harrison et al. 2020). Also, we note that if a cellular ISD is used for metabarcoding studies it is wise to consider the CNV of the focal loci when performing concentration calculations prior to ISD addition (see Stämmler et al. 2016).

ISDs are not a panacea for all the ills of sequencing

ISDs can account for among-sample variation when comparing the effects of treatment or ecological covariates on abundances (both relative and absolute) of a particular taxon. They cannot however address all the concerns that complicate the comparison of abundances of *different* taxa among and within samples. In part, this is because no ISD, or mixture of ISDs, is perfectly commutable with each taxon in a complex ecological assemblage. It must be remembered that every step of the library preparation process has the potential to impose idiosyncratic, selective biases for and against the DNA sequences associated with different taxa in a sample (Fig. 3; Nilsson et al. 2018). For example, PCR primers do not match their target sequences equally well in all taxa, leading to preferential amplification of some taxa, and substantial differences in selectivity among different primers (Fouhy et al. 2016; Hong et al. 2009). Thus, if a primer pair is biased against a particular sequence, then the abundance within the sample will be underestimated and an ISD cannot remedy this error. Primer bias is a well known issue, but nearly every other aspect of PCR can also impose unwanted biases—including the type of polymerase and reagents used (Nilsson et al. 2018; Pollock et al. 2018; Schori et al. 2013), cycle count (Kelly et al. 2019; Silverman et al. 2019), GC content (Laursen et al. 2017; Risso et al. 2011), and length of the amplicon (Oshlack and Wakefield 2009). Aside from PCR, even the choice of sequencing platform can impose bias (D'Amore et al. 2016). Thus, these procedural biases can cause false negatives in inferences about external determinants of assemblage composition and simply make it difficult to know true abundances.

Estimates of abundances of taxa are further complicated by error that arises due to high copy number variation (CNV) among taxa in marker loci. For example, Lofgren et al. (2019) reported that fungal taxa can differ in ITS copy number by an order of magnitude or more. Even within a single fungal taxon, *Suillus brevipes*, ITS copy number ranged from 72–156. While not quite as extreme as for fungi, CNV is also widespread among bacteria for the commonly used 16S marker (Kembel et al. 2012; Lee et al. 2009; Perisin et al. 2016;

Stoddard et al. 2015; Větrovský and Baldrian 2013). Of course, variation in ploidy-level (Pecoraro et al. 2011), or the number of nuclei in a cell (which can vary for some multicellular fungi; Gladieux et al. 2014), can also influence copy number variation. A possible mitigation solution for bacteria and archaea is bioinformatic correction of CNV of focal taxa via comparison to the popular rrnDB database (Stoddard et al. 2015).

A special, but similar, problem exists for researchers studying environmental DNA to characterize assemblages of multi-cellular organisms, as taxa shed different numbers of cells (e.g., due to variation in body size or in germ cell production) and live for different amounts of time (Thomsen and Willerslev 2015). Thus one individual of an organism could, over its lifetime, shed many more cells than multiple individuals of organisms with different traits (Cristescu and Hebert 2018).

When taken together, these biases suggest extreme caution is in order when interpreting sequence data with the intention of inter-taxa comparisons of abundance (Fig. 3), such as when analyses focus on description of overall shifts in community composition as defined by changes in rank order abundances among taxa. Unfortunately, many ecology studies rely on a common suite of such analyses, including description of patterns in diversity entropies, ordination techniques, and PERMANOVA. If taxon-specific analyses are used instead, or in conjunction with these techniques, many of the biases we describe here become much less problematic. This is because most biases will affect a taxon in the same way across samples and, therefore, biases will not be confounded with experimental treatment(s) or ecological covariates of interest (McLaren et al. 2019; Morton et al. 2019). Moreover, many ecological questions are better answered by quantifying the effect of treatment on specific taxa, rather than documenting shifts in overall assemblage composition. We note that if inter-taxon analyses are required, that conversion to absolute abundances removes at least some of the challenges imposed by compositionality that confound such inferences. Indeed, a primary benefit of ISDs are that they allow many popular community ecology statistics to be employed—many statistical techniques are inappropriate for compositional data (Gloor et al. 2017; Jackson 1997).

To learn about the biological causes of differences in taxon abundances among samples, it is helpful to partition variation that arises from replicated laboratory processes and biological variation among samples. Assuming commutability of ISDs, technical variation among replicates can be estimated for the ISDs and subtracted from variation for individual taxa to yield an estimate of biological variation for each taxon (Ji et al. 2019; Risso et al. 2014). Ji et al. (2019) recently used such an approach to isolate spatial, temporal, and technical variation in absolute abundances of gut microbes. The bulk of the variation they observed was assigned to technical causes. We suggest that Bayesian models are an exciting possibility for partitioning variation in sequence data, in part because they make full use of the data and can incorporate hierarchical model structures to share information among all replicates within a sampling group (*sensu* Fordyce et al. 2011; Harrison et al. 2020). This is in contrast to rarefaction methods, which discard observed data and thus provide potentially misleading information about technical and biological variation among samples (McMurdie and Holmes 2014).

Conclusion

Sequencing is a powerful tool to measure abundance of organisms that are difficult to observe and count directly. As a research community, we are growing increasingly aware of the drawbacks of compositional sequencing data and the benefits provided by ISDs. But, as we have shown, the efficacy of ISDs is dependent upon careful accounting during laboratory practices and potentially unrealistic assumptions of biological simplicity (e.g., in CNV). Notwithstanding these challenges, ISDs liberate researchers from the constraints imposed by relative abundance data and we suggest that their use become a standard component of sequence-based study of ecological assemblages.

Acknowledgements

This research was supported by the Microbial Ecology Collaborative at the University of Wyoming with funding from NSF award #EPS-1655726. We wish to thank Gregory Randolph for helpful conversations and three anonymous reviewers for constructive comments on an earlier version of this manuscript.

Data Accessibility

There are no data associated with this publication.

References

Aitchison, J. (1982). *The statistical analysis of compositional data*. New York, NY: Chapman and Hall.

Aitchison, J. and J. J. Egozcue (2005). “Compositional data analysis: where are we and where should we be heading?” *Mathematical Geology* 37.7, pp. 829–850.

Amann, R. and B. M. Fuchs (2008). “Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques”. *Nature Reviews Microbiology* 6.5, pp. 339–348.

Baker, M. (2012). “Digital PCR hits its stride”. *Nature Methods* 9.6, pp. 541–544.

Barlow, J. T., S. R. Bogatyrev, and R. F. Ismagilov (2020). “A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities”. *Nature Communications* 11.1, p. 2590.

Benninghoff, W. S. (1962). “Calculation of pollen and spore density in sediments by addition of exotic pollen in known quantities”. *Pollen et Spores* 4, pp. 332–333.

Bonk, F. et al. (2018). “PCR-based quantification of taxa-specific abundances in microbial communities: Quantifying and avoiding common pitfalls”. *Journal of Microbiological Methods* 153, pp. 139–147.

Caporaso, J. G. et al. (2012). “Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms”. *The ISME Journal* 6.8, pp. 1621–1624.

Carini, P. (2019). “A “cultural” renaissance: genomics breathes new life into an old craft”. *mSystems* 4.3, e00092–19.

Chen, K. et al. (2016). “The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses”. *Molecular and Cellular Biology* 36.5, pp. 662–667.

Cristescu, M. E. and P. D. Hebert (2018). “Uses and misuses of environmental DNA in biodiversity science and conservation”. *Annual Review of Ecology, Evolution, and Systematics* 49.1, pp. 209–230.

D’Amore, R. et al. (2016). “A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling”. *BMC Genomics* 17.

Daims, H. et al. (2001). “Cultivation-independent, semiautomatic determination of absolute bacterial cell numbers in environmental samples by fluorescence in situ hybridization”. *Applied and Environmental Microbiology* 67.12, pp. 5810–5818.

Dannemiller, K. C. et al. (2014). “Combining real-time PCR and next-generation DNA sequencing to provide quantitative comparisons of fungal aerosol populations”. *Atmospheric Environment* 84, pp. 113–121.

Davis, M. B. (1963). “On the theory of pollen analysis”. *American Journal of Science* 261.10, pp. 897–912.

— (1966). “Determination of absolute pollen frequency”. *Ecology* 47.2, pp. 310–311.

Davis, M. B. and E. S. Deevey (1964). “Pollen accumulation rates: estimates from late-glacial sediment of Rogers Lake”. *Science* 145.3638, pp. 1293–1295.

Deagle, B. E. et al. (2018). “Genetic monitoring of open ocean biodiversity: An evaluation of DNA metabarcoding for processing continuous plankton recorder samples”. *Molecular Ecology Resources* 18.3, pp. 391–406.

Doležel, J., J. Greilhuber, and J. Suda (2007). “Estimation of nuclear DNA content in plants using flow cytometry”. *Nature Protocols* 2.9, pp. 2233–2244.

Egozcue, J. J. et al. (2003). “Isometric logratio transformations for compositional data analysis”. *Mathematical Geology* 35.3, pp. 279–300.

Eisenberg, E. and E. Y. Levanon (2013). “Human housekeeping genes, revisited”. *Trends in Genetics*. Human Genetics 29.10, pp. 569–574.

Fagerlind, F. (1952). “The real signification of pollen diagrams”. *Botaniska Notiser* 105, pp. 185–224.

Fernandes, A. D. et al. (2014). “Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis”. *Microbiome* 2, p. 15.

Flores, G. E., J. B. Henley, and N. Fierer (2012). “A direct PCR approach to accelerate analyses of human-associated microbial communities”. *PLoS ONE* 7.9.

Fordyce, J. A. et al. (2011). “A hierarchical Bayesian approach to ecological count data: a flexible tool for ecologists”. *PLOS ONE* 6.11, e26785.

Fouhy, F. et al. (2016). “16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform”. *BMC Microbiology* 16.1, p. 123.

Frossard, A., F. Hammes, and M. O. Gessner (2016). “Flow cytometric assessment of bacterial abundance in soils, sediments and sludge”. *Frontiers in Microbiology* 7.

Giesecke, T. and S. L. Fontana (2008). “Revisiting pollen accumulation rates from Swedish lake sediments”. *The Holocene* 18.2, pp. 293–305.

Gladieux, P. et al. (2014). “Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes”. *Molecular Ecology* 23.4, pp. 753–773.

Gloor, G. B. and G. Reid (2016). “Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data”. *Canadian Journal of Microbiology* 62.8, pp. 692–703.

Gloor, G. B. et al. (2017). “Microbiome datasets are compositional: and this is not optional”. *Frontiers in Microbiology* 8.

Goodrich, J. K. et al. (2014). “Conducting a microbiome study”. *Cell* 158.2, pp. 250–262.

Guo, X. et al. (2019). “Host-associated quantitative abundance profiling reveals the microbial load variation of root microbiome”. *Plant Communications*, p. 100003.

Hardwick, S. A., I. W. Deveson, and T. R. Mercer (2017). “Reference standards for next-generation sequencing”. *Nature Reviews Genetics* 18.8, pp. 473–484.

Hardwick, S. A. et al. (2016). “Spliced synthetic genes as internal controls in RNA sequencing experiments”. *Nature Methods* 13.9, pp. 792–798.

Hardwick, S. A. et al. (2018). “Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis”. *Nature Communications* 9.1, pp. 1–10.

Harrison, J. G. et al. (2020). “Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data”. *Molecular Ecology Resources* 20.2, pp. 481–497.

Higuchi, R. et al. (1993). “Kinetic PCR analysis: real-time monitoring of DNA amplification reactions”. *Bio/Technology* 11.9, pp. 1026–1030.

Hindson, B. J. et al. (2011). “High-throughput droplet digital PCR system for absolute quantitation of DNA copy number”. *Analytical Chemistry* 83.22, pp. 8604–8610.

Hong, S. et al. (2009). “Polymerase chain reaction primers miss half of rRNA microbial diversity”. *The ISME Journal* 3.12, pp. 1365–1373.

Hossain, A. et al. (2020). *A massively parallel COVID-19 diagnostic assay for simultaneous testing of 19200 patient samples*. Google Docs. URL: `https://docs.google.com/document/d/1kP2w_uTMSep2UxTCOnUhh1TMCjWvHEYOsUUpkJHPYV4/preview?sle=true&usp=embed_facebook` (visited on 2020).

Huggett, J. F. et al. (2008). “Differential susceptibility of PCR reactions to inhibitors: an important and unrecognised phenomenon”. *BMC Research Notes* 1.1, p. 70.

Humphrey, P. T. and N. K. Whiteman (2020). “Insect herbivory reshapes a native leaf microbiome”. *Nature Ecology & Evolution* 4.2, pp. 221–229.

Jackson, D. A. (1997). “Compositional data in community ecology: the paradigm or peril of proportions?” *Ecology* 78.3, pp. 929–940.

Ji, B. W. et al. (2019). “Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling”. *Nature Methods* 16.8, pp. 731–736.

Ji, Y. et al. (2020). “SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes”. *Molecular Ecology Resources* 20.1, pp. 256–267.

Jian, C. et al. (2020). “Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling”. *PLOS ONE* 15.1, e0227285.

Jiang, L. et al. (2011). “Synthetic spike-in standards for RNA-seq experiments”. *Genome Research.*

Jones, M. B. et al. (2015). “Library preparation methodology can influence genomic and functional predictions in human microbiome research”. *Proceedings of the National Academy of Sciences* 112.45, pp. 14024–14029.

Jonge, H. J. M. de et al. (2007). “Evidence based selection of housekeeping genes”. *PLoS ONE* 2.9.

Karasov, T. L. et al. (2018). “*Arabidopsis thaliana* and *Pseudomonas* pathogens exhibit stable associations over evolutionary timescales”. *Cell Host & Microbe* 24.1, 168–179.e4.

Karasov, T. L. et al. (2019). “The relationship between microbial biomass and disease in the *Arabidopsis thaliana* phyllosphere”. *bioRxiv*, p. 828814.

Kelly, R. P., A. O. Shelton, and R. Gallego (2019). “Understanding PCR processes to draw meaningful conclusions from environmental DNA studies”. *Scientific Reports* 9.1, p. 12133.

Kembel, S. W. et al. (2012). "Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance". *PLOS Computational Biology* 8.10, e1002743.

Kim, T. G., S.-Y. Jeong, and K.-S. Cho (2015). "Development of droplet digital PCR assays for methanogenic taxa and examination of methanogen communities in full-scale anaerobic digesters". *Applied Microbiology and Biotechnology* 99.1, pp. 445–458.

Langille, M. G. I. et al. (2013). "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences". *Nature Biotechnology* 31.9, pp. 814–821.

Laursen, M. F., M. D. Dalgaard, and M. I. Bahl (2017). "Genomic GC-content affects the accuracy of 16S rRNA gene sequencing based microbial profiling due to PCR bias". *Frontiers in Microbiology* 8.

Lee, Z. M.-P., C. Bussema, and T. M. Schmidt (2009). "rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea". *Nucleic Acids Research* 37 (suppl_1), pp. D489–D493.

Lin, Y. et al. (2019). "Towards quantitative microbiome community profiling using internal standards". *Applied and Environmental Microbiology* 85.5.

Lofgren, L. A. et al. (2019). "Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles". *Molecular Ecology* 28.4, pp. 721–730.

Lou, J. et al. (2018). "Assessing soil bacterial community and dynamics by integrated high-throughput absolute abundance quantification". *PeerJ* 6, e4514.

Lun, A. T. L. et al. (2017). "Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data". *Genome Research* 27.11, pp. 1795–1806.

Mahmoudi, N., G. F. Slater, and R. R. Fulthorpe (2011). "Comparison of commercial DNA extraction kits for isolation and purification of bacterial and eukaryotic DNA from PAH-contaminated soils". *Canadian Journal of Microbiology* 57.8, pp. 623–628.

McLaren, M. R., A. D. Willis, and B. J. Callahan (2019). "Consistent and correctable bias in metagenomic sequencing measurements". *bioRxiv*, p. 559831.

McMurdie, P. J. and S. Holmes (2014). "Waste not, want not: why rarefying microbiome data Is inadmissible". *PLOS Comput Biol* 10.4, e1003531.

Morella, N. M. et al. (2018). "Rapid quantification of bacteriophages and their bacterial hosts in vitro and in vivo using droplet digital PCR". *Journal of Virological Methods* 259, pp. 18–24.

Morton, J. T. et al. (2019). "Establishing microbial composition measurement standards with reference frames". *Nature Communications* 10.1, p. 2719.

Nguyen, N. H. et al. (2014). "Parsing ecological signal from noise in next generation amplicon sequencing". *New Phytologist* 205.4, pp. 1389–1393.

Nilsson, R. H. et al. (2018). "Mycobiome diversity: high-throughput sequencing and identification of fungi". *Nature Reviews Microbiology*, p. 1.

Opel, K. L., D. Chung, and B. R. McCord (2010). "A study of PCR inhibition mechanisms using real time PCR". *Journal of Forensic Sciences* 55.1, pp. 25–33.

Oshlack, A. and M. J. Wakefield (2009). "Transcript length bias in RNA-seq data confounds systems biology". *Biology Direct* 4.1, p. 14.

Pearson, K. (1897). "Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurement of organs". *Proceedings of the Royal Society of London* 60.359, pp. 489–498.

Pecoraro, V. et al. (2011). “Quantification of ploidy in Proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species”. *PLoS ONE* 6.1.

Perisin, M. et al. (2016). “16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies”. *The ISME Journal* 10.4, pp. 1020–1024.

Piwosz, K. et al. (2018). “Determining lineage-specific bacterial growth curves with a novel approach based on amplicon reads normalization using internal standard (ARNIS)”. *The ISME Journal* 12.11, pp. 2640–2654.

Pollock, J. et al. (2018). “The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies”. *Applied and Environmental Microbiology* 84.7, e02627–17.

Prentice, I. C. and T. Webb (1986). “Pollen percentages, tree abundances and the Fagerlind effect”. *Journal of Quaternary Science* 1.1, pp. 35–43.

Props, R. et al. (2017a). “Absolute quantification of microbial taxon abundances”. *The ISME journal* 11.2, pp. 584–587.

Props, R. et al. (2017b). “Measuring the biodiversity of microbial communities by flow cytometry”. *Methods in Ecology and Evolution*, pp. 1376–1385.

Quinn, T. P. et al. (2018). “Understanding sequencing data as compositions: an outlook and review”. *Bioinformatics* 34.16, pp. 2870–2878.

Quinn, T. P. et al. (2019). “A field guide for the compositional analysis of any-omics data”. *GigaScience* 8.9.

Regalado, J. et al. (2019). “Combining whole genome shotgun sequencing and rDNA amplicon analyses to improve detection of microbe-microbe interaction networks in plant leaves”. *bioRxiv*, p. 823492.

Risso, D. et al. (2011). “GC-content normalization for RNA-Seq data”. *BMC Bioinformatics* 12.1, p. 480.

Risso, D. et al. (2014). “Normalization of RNA-seq data using factor analysis of control genes or samples”. *Nature Biotechnology* 32.9, pp. 896–902.

Rossen, L. et al. (1992). “Inhibition of PCR by components of food samples, microbial diagnostic assays and DNA-extraction solutions”. *International Journal of Food Microbiology* 17.1, pp. 37–45.

Schmidt, P.-A. et al. (2013). “Illumina metabarcoding of a soil fungal community”. *Soil Biology and Biochemistry* 65, pp. 128–132.

Schori, M. et al. (2013). “Engineered DNA polymerase improves PCR results for plastid DNA”. *Applications in Plant Sciences* 1.2, p. 1200519.

Schrader, C. et al. (2012). “PCR inhibitors – occurrence, properties and removal”. *Journal of Applied Microbiology* 113.5, pp. 1014–1026.

Silverman, J. D. et al. (2019). “Measuring and mitigating PCR bias in microbiome data”. *bioRxiv*, p. 604025.

Smets, W. et al. (2016). “A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing”. *Soil Biology and Biochemistry* 96, pp. 145–151.

Stämmler, F. et al. (2016). “Adjusting microbiome profiles for differences in microbial load by spike-in bacteria”. *Microbiome* 4.1, p. 28.

Stoddard, S. F. et al. (2015). “rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development”. *Nucleic Acids Research* 43 (D1), pp. D593–D598.

Taberlet, P. et al. (2012). “Towards next-generation biodiversity assessment using DNA metabarcoding”. *Molecular Ecology* 21.8, pp. 2045–2050.

Thellin, O. et al. (1999). “Housekeeping genes as internal standards: use and limits”. *Journal of Biotechnology* 75.2, pp. 291–295.

Thomsen, P. F. and E. Willerslev (2015). “Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity”. *Biological Conservation.* Special Issue: Environmental DNA: A powerful new tool for biological conservation 183, pp. 4–18.

Tkacz, A., M. Hortala, and P. S. Poole (2018). “Absolute quantitation of microbiota abundance in environmental samples”. *Microbiome* 6.1, p. 110.

Tourlousse, D. M., A. Ohashi, and Y. Sekiguchi (2018). “Sample tracking in microbiome community profiling assays using synthetic 16S rRNA gene spike-in controls”. *Scientific Reports* 8.1, pp. 1–9.

Tourlousse, D. M. et al. (2017). “Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing”. *Nucleic Acids Research* 45.4, e23–e23.

Tricarico, C. et al. (2002). “Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies”. *Analytical Biochemistry* 309.2, pp. 293–300.

Tsilimigras, M. C. B. and A. A. Fodor (2016). “Compositional data analysis of the microbiome: fundamentals, tools, and challenges”. *Annals of Epidemiology.* The Microbiome and Epidemiology 26.5, pp. 330–335.

Turner, C. R., K. L. Uy, and R. C. Everhart (2015). “Fish environmental DNA is more concentrated in aquatic sediments than surface water”. *Biological Conservation.* Special Issue: Environmental DNA: A powerful new tool for biological conservation 183, pp. 93–102.

Vandeputte, D. et al. (2017). “Quantitative microbiome profiling links gut community variation to microbial load”. *Nature* 551.7681, pp. 507–511.

Venkataraman, A. et al. (2018). “Spike-in genomic DNA for validating performance of metagenomics workflows”. *BioTechniques* 65.6, pp. 315–321.

Větrovský, T. and P. Baldrian (2013). “The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses”. *PLOS ONE* 8.2, e57923.

Videvall, E. et al. (2017). “Direct PCR offers a fast and reliable alternative to conventional DNA isolation methods for gut microbiomes”. *mSystems* 2.6.

Weiss, S. et al. (2017). “Normalization and microbial differential abundance strategies depend upon data characteristics”. *Microbiome* 5, p. 27.

Wilson, D. and G. C. Carroll (1997). “Avoidance of high-endophyte space by gall-forming insects”. *Ecology* 78.7, pp. 2153–2163.

Wilson, I. G. (1997). “Inhibition and facilitation of nucleic acid amplification.” *Applied and Environmental Microbiology* 63.10, pp. 3741–3751.

Yang, L. et al. (2018). “Use of an improved high-throughput absolute abundance quantification method to characterize soil bacterial community and dynamics”. *Science of The Total Environment* 633, pp. 360–371.

Zemb, O. et al. (2020). “Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard”. *MicrobiologyOpen*, e977.

Zhang, Z. et al. (2017). “Soil bacterial quantification approaches coupling with relative abundances reflecting the changes of taxa”. *Scientific Reports* 7.1, pp. 1–11.

Table 1: Publications that describe development of internal standards (ISDs). Publications are organized by the type of ISD—either cellular, or biological or synthetic DNA. Notes regarding the design and suggested usage of the ISDs are mentioned. We use ‘synthetic’ to refer to DNA sequences designed to avoid similarity to known biological sequences. We sought to represent the variety in ISD designs here, not every instance of their use. The body of literature surrounding ISDs is rapidly growing and we apologize to any authors that we have inadvertently omitted from this list. See the main text for discussion of non-ISD approaches to absolutely quantify sequenced taxa (e.g., via qPCR, flow cytometery, or through comparison to host-derived reads).

Citation	**Design (taxon or sequence)**	**Usage**
	Cellular ISDs	
K. Piwosz et al. (2018). “Determining lineage-specific bacterial growth curves with a novel approach based on amplicon reads normalization using internal standard (ARNIS)”. *The ISME Journal* 12.11, pp. 2640–2654	*Escherichia coli* (a model bacterium)	Added to water samples prior to extraction. Target concentration of ISD was 5% of the bacterial assemblage.
F. Stämmler et al. (2016). “Adjusting microbiome profiles for differences in microbial load by spike-in bacteria”. *Microbiome* 4.1, p. 28	*Salinibacter ruber* (halophilic), *Rhizobium radiobacter* (rhizosphere inhabitant) and *Alicyclobacillus acidiphilus* (thermoacidophile)	Added before extraction of stool samples in proportion to 16S CNV differences among ISD taxa. Reported that combining data from ISDs reduced error.
B. W. Ji et al. (2019). “Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling”. *Nature Methods* 16.8, pp. 731–736	*Sporosarcina pasteurii* (an environmental bacterium not present in focal samples)	Added ISD before extraction of stool samples. This publication contains a novel way to partition spatial, temporal, and technical variation using ISDs.

M. B. Jones et al. (2015). “Library preparation methodology can influence genomic and functional predictions in human microbiome research”. *Proceedings of the National Academy of Sciences* 112.45, pp. 14024–14029	*Shewanella oneidensis* (a soil and marine sediment bacterium)	Early use of cellular ISD. Used to demonstrate technical variation of sequencing and qPCR of a mock community and stool samples.
Genomic or amplified biological DNA ISDs		
B. E. Deagle et al. (2018). “Genetic monitoring of open ocean biodiversity: An evaluation of DNA metabarcoding for processing continuous plankton recorder samples”. *Molecular Ecology Resources* 18.3, pp. 391–406	Mouse genomic DNA	Added to plankton samples prior to extraction. 225 ng of ISD was added per ~50 mg of sample. This resulted in 0.5–66% of sequences in each sample. ISD useful for detection of poor samples.
Y. Ji et al. (2020). “SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes”. *Molecular Ecology Resources* 20.1, pp. 256–267	Amplified and barcoded COI (658 bp) from *Bombyx mori* and two unnamed beetle taxa	Added the three ISDs before extraction of bulk invertebrate samples in a 1:2:4 ratio.
Y. Lin et al. (2019). “Towards quantitative microbiome community profiling using internal standards”. *Applied and Environmental Microbiology* 85.5	Genomic DNA from *Schizosaccharomyces pombe* (model yeast) for 18S and *Thermus thermophilus* (model thermophilic bacterium) for 16S	Added to seawater samples prior to extraction. Aimed for ISDs capturing 1% of reads and were close to this ideal.
W. Smets et al. (2016). “A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing”. *Soil Biology and Biochemistry* 96, pp. 145–151	Genomic DNA from *Aliivibrio fischeri* (model symbiotic bacterium of squid) and *Thermus thermophilus* (model thermophilic bacterium)	ISDs added to soil samples prior to DNA extraction. Tried using ISD concentrations of 0.1 or 1% of the expected yield of DNA.

A. Venkataraman et al. (2018). “Spike-in genomic DNA for validating performance of metagenomics workflows”. *BioTechniques* 65.6, pp. 315–321	Genomic DNA of *Alivibrio fischeri* (model symbiotic bacterium of squid) and *Rhodopseudomonas palustris* (marine bacterium)	Spiked into samples in a 4:1 ratio before DNA extraction. ISDs suggested for use in shotgun metagenomics.
Synthetic DNA ISDs		
S. A. Hardwick et al. (2018). “Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis”. *Nature Communications* 9.1, pp. 1–10	86 synthetic sequences modeled after diverse taxa and ~1–10 kb long. Sequences are reversed portions of focal genomes.	Suitable for metagenomics, including via long-read technology. Adaptation for metabarcoding would require primer sequence addition.
L. Jiang et al. (2011). “Synthetic spike-in standards for RNA-seq experiments”. *Genome Research*	96 synthetic RNAs, spanning a variety of lengths, GC contents, and 2^{20} range in concentrations	Designed to be used as a mixture of ISDs for transcriptomic studies.
D. M. Tourlousse et al. (2017). “Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing”. *Nucleic Acids Research* 45.4, e23–e23	12 ISDs composed of conserved biological sequences interspersed with synthetic sequences. Care was taken to ensure balanced GC content, no homopolymers over 3 bp, limited repeat density and self-incompatibility of synthetic regions. Sequences are ~1,500 bp long.	ISDs were added to soil samples at the point of extraction, after cell lysis
A. Tkacz et al. (2018). “Absolute quantitation of microbiota abundance in environmental samples”. *Microbiome* 6.1, p. 110	Three synthetic ISDs with primer regions for either 16S, 18S, or ITS1. GC content of the synthetic region was designed to match that of focal taxa.	Added both before and after DNA extraction and reported superior performance for the former.

O. Zemb et al. (2020). “Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard”. *MicrobiologyOpen*, e977	Amplified *Escherichia coli* rRNA (733 bp) with synthetic regions interjected.	Suggested a combination of qPCR and sequencing.

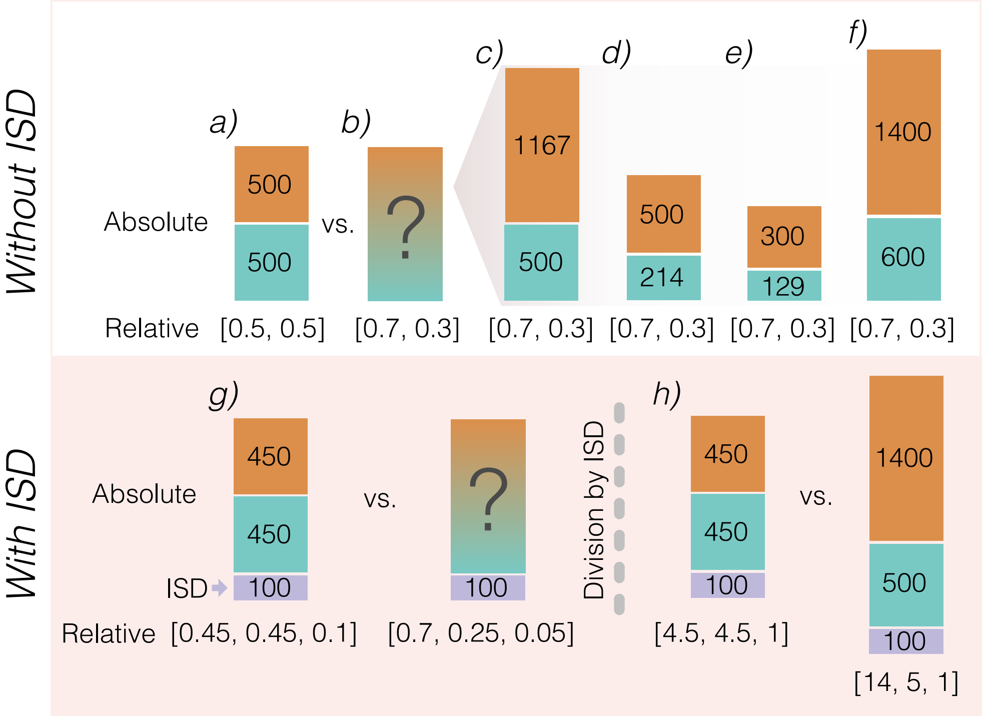

Figure 1: The problem of compositionality and how an internal standard (ISD) can help. Panels a and b show the absolute and relative abundances of two hypothetical samples that are each representatives of differing experimental conditions—say from a treatment-control experimental design. Each sample contains two taxa (shown in blue and orange respectively). Panels c–f demonstrate the many different absolute abundances for sample b that could give rise to the same relative abundance profile. One taxon could increase (c); or decrease (d); or both taxa could decrease, but one more so than the other (e); or both taxa could increase, but one more so than the other (f). Thus, it is not possible to determine shifts in absolute abundances from relative abundance data. However, if a consistent amount of an ISD is added to each sample (panel g), then division by the ISD (panel h) can convert relative abundance data into ratios that are proportional to the absolute abundances present in each sample. Estimation of absolute abundances is possible upon multiplication of proportions by a constant that encompasses variation in extracted mass while accounting for copy-number variation (if appropriate, see main text).

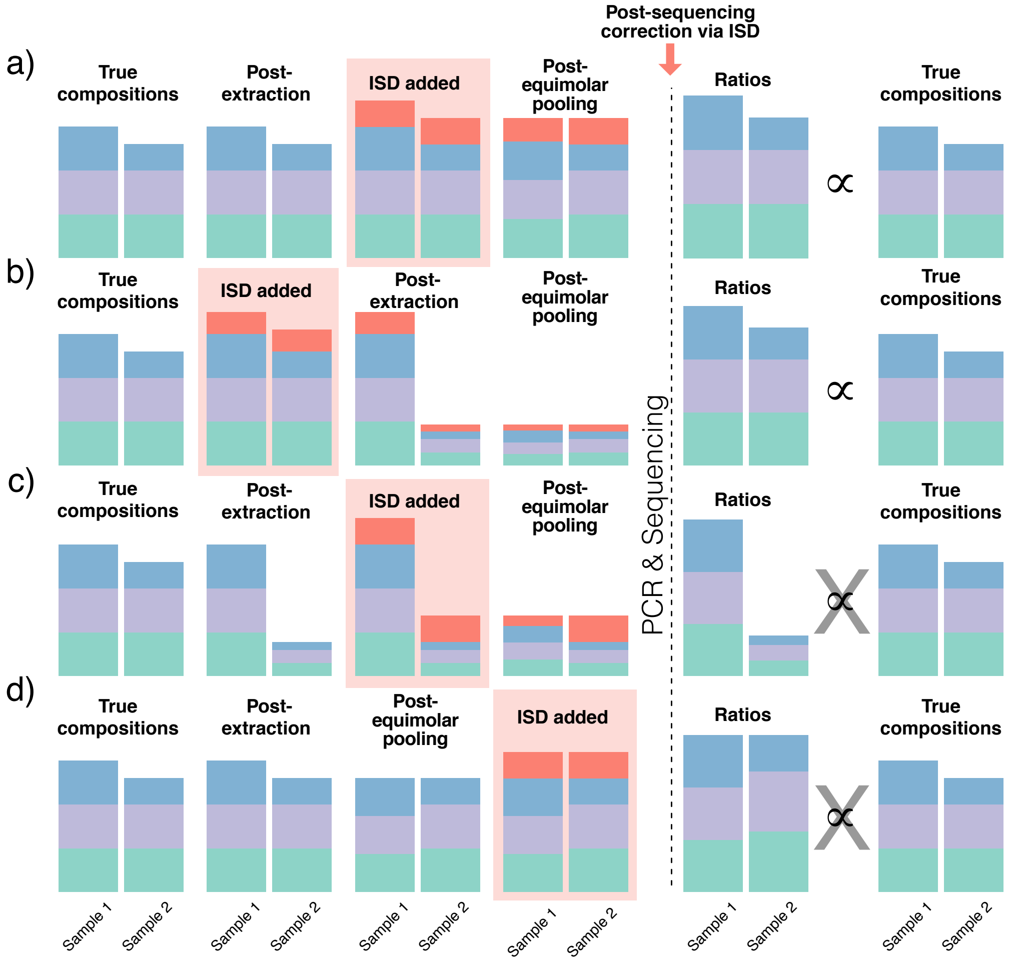

Figure 2: The addition of an internal standard (ISD) to samples can correct for the problems posed by the compositional nature of sequencing data, but the ISD must be added at the correct time during sample processing. Here, we present data representative of four laboratory scenarios that affect ISD efficacy. For each scenario, we present relative abundance data for two samples, each of which contains three features that are shown in different colors. The ISD is shown in orange and, for each scenario, a light orange box denotes the step at which the ISD is added. a) Here, ISD is added prior to equimolar pooling of nucleic acids (a common practice prior to PCR and sequencing) and there is no variation in sample mass, or yield from nucleic acid extraction, or other biases induced by laboratory-practice. In this case the ISD performs as desired. b) If the ISD is added prior to nucleic acid extraction and reflects variation in extraction yield among samples (i.e., as would be expected for a commutable cellular ISD), then the ISD can be used to back-calculate absolute abundances. c) However, if the ISD is added after extraction, or is not commutable to focal taxa during extraction, and samples differ in extraction yield, then the ISD will not perform as expected. d) Similarly, if the ISD is added after equimolar pooling of samples then it is no longer effective.

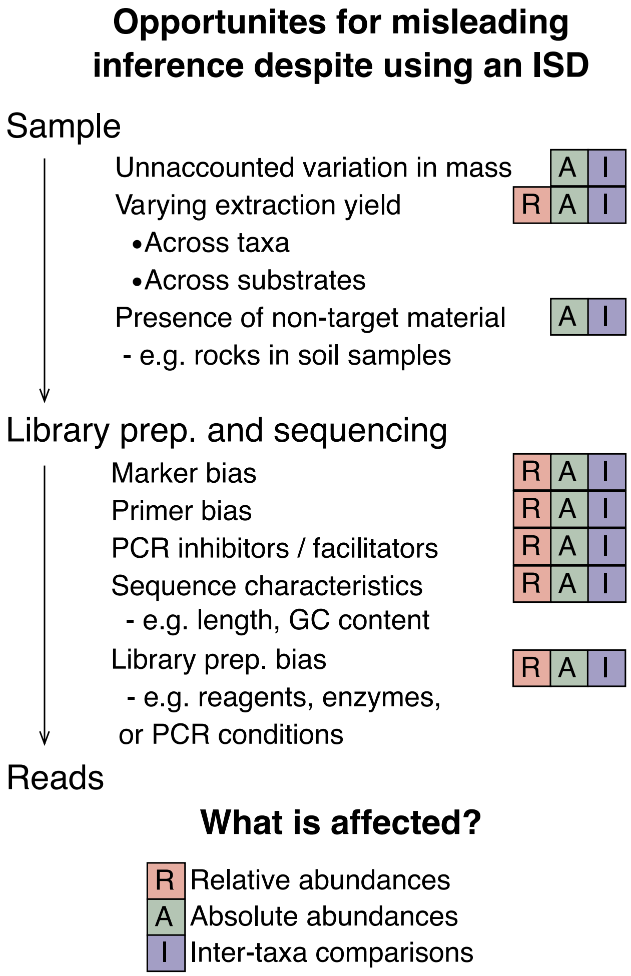
Opportunites for misleading inference despite using an ISD
Sample
Unnaccounted variation in mass
A I
Varying extraction yield
R A I
•Across taxa
•Across substrates
Presence of non-target material
A I
- e.g. rocks in soil samples
Library prep. and sequencing
Marker bias
R A I
Primer bias
R A I
PCR inhibitors / facilitators
R A I
Sequence characteristics
R A I
- e.g. length, GC content
Library prep. bias
R A I
- e.g. reagents, enzymes, or PCR conditions
Reads
What is affected?
R Relative abundances
A Absolute abundances
I Inter-taxa comparisons

Figure 3: ISDs are useful tools, but cannot correct for all biases associated with sequence-based characterization of ecological assemblages. We present here a selection of biases that are organized chronologically following the data generation process—from sampling to sequencing. Colored boxes next to each source of bias denote whether it can affect relative abundances or absolute abundances. It is likely that many of the biases mentioned here act in a taxon-specific manner, thus inter-taxa comparisons of abundance are fraught (i.e., comparing taxa in terms of cell count within or among samples or sampling groups). Biases can affect many analyses. For instance, differential abundance analysis among sampling groups of relative or absolute abundance data will be misleading if biases affect accurate estimation of either type of abundance. This catalogue of biases does not mean sequence-based characterization of ecological assemblages is doomed to fail, only that biases must be carefully considered when planning an experiment so that the most meaning can be extracted from the resulting data.